NDA/Demonstrating Product Effectiveness

Robert J. Temple, M.D.
Associate Director for Medical Policy
Center for Drug Evaluation and Research
U.S. Food and Drug Administration

Unapproved Drugs Workshop
January 9, 2007

Demonstrating Effectiveness

I will discuss the "harder" cases, where effectiveness is not established by:

- DESI effective rating
- Approved drug NDA
- Approved NDA or DESI combination containing the drug [we concluded that each component was effective]

In those cases bioavailability and chemistry are generally all that's needed for the same drug and possibly even for a different salt or ester (which, technically is a different drug but the same active moiety).

If the dosage form is different, studies may be needed (not for tablet/capsule; maybe for controlled release; certainly for most changes in route-inhaled, topical, but perhaps not all, such as injection "tide-over")

Demonstrating Effectiveness

If effectiveness of the active moiety is not established, approval requires that it be established. Generally the route for doing this is the NDA, whose effectiveness standard I will discuss.

Monographs (for OTC drugs) or seeking a determination of GRAE do not represent an escape. Effectiveness is established for drugs in a monograph more or less identically to NDA drugs.

GRAE is, if anything, a higher standard [Weinberger vs Hynson, Westcott, and Dunning: a consensus among experts. . . Based on published scientific literature of the same quantity and quality needed to approve a drug under section 505 of the Act].

Legal Standard

“New Drugs” must be shown effective under 505 (d)(5):

“substantial evidence that the drug will have the effect it purports or is represented to have under the conditions of use prescribed, recommended, or suggested in the proposed labeling.”

“substantial evidence means evidence consisting of adequate and well-controlled investigations. . . By [qualified] aspects. . . on the basis of which it could fairly and responsibly be concluded by such experts that the drug will have the effect [represented in labeling].”

Note:

1. The interpreting experts are FDA
2. The effect has to be meaningful

[Warner-Lambert v Heckler, 1986]

Legal Standard

The plural in investigations was intended. FDAMA allows reliance on a single study plus "confirmatory evidence" but for symptomatic conditions it would be unusual for us to accept a single study. But the studies don't need to be identical and diverse sorts of data can provide support [Guidance: Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products, 1998]

Legal Standard

The requirement is thus twofold:

- The supportive studies need to be "well-controlled"
- They need to be convincing

As a historical matter, two studies showing well-controlled, properly analyzed "statistical significance" (a 2-sided p-value of < 0.05) have been considered to be convincing to experts.

We have sometimes relied on a single stronger study, ($p = 0.01 – 0.001$) but usually for important outcomes.

Adequate and Well-Controlled Studies

21 CFR 314.126 gives the characteristics of an A&WC study. Briefly, they are

1. Comparison of the treatment with a control

   Because the course of most diseases, is variable, you need a control group, a group treated just like the test group, <u>except</u> that they don't get the drug, to distinguish the effect of the drug from spontaneous change, placebo effect, observer expectations.

Adequate and Well-Controlled Studies

1. Control (cont)

The rule describes 5 kinds of control

- Placebo
- No treatment
- Dose response
- Active – superiority or Non-Inferiority
- Historical

For symptomatic conditions, randomization and blinding are needed and NI or historically controlled trials are unlikely to be persuasive.

Therefore, placebo or dose-response are the usual designs needed.

Adequate and Well-Controlled Studies

2. Minimization of bias: a "tilt" favoring one group, a directed (non-random) difference in how test and control group are selected, treated, observed, and analyzed (the 4 main places bias can enter).

   Remedies
   - Blinding (patient and observer bias)
   - Randomization (treatment and control start out equal)
   - Careful specification of procedures and analyzes in a protocol to avoid
     - Choosing the most favorable analysis out of many (bias)
     - Having so many analyses that one is favorable by chance (multiplicity)

Adequate and Well-Controlled Studies

3. Sufficient detail to know how the study was done and what the results were

   This was a major problem in the past and is definitely a problem if one is trying to rely on old literature. In those cases (still true today), analytic plan is rarely specified, handling of dropouts is rarely described, other therapy is not discussed. It is sometimes hard to tell duration of treatment and other critical details.

Adequate and Well-Controlled Studies

The basic principles were described in a 1970 rule, updated 1985, but we've learned a great deal, often from the DESI experience:

Just a few illustrations:

1. Interim looks at data
2. Counting all patients
3. Changing analyses
4. Active control non-inferiority trials
5. Having all the details

Interim Looks

If you monitor results as they come in, and stop when a goal is attained, you are likely to see "an effect" at some point, because of random variation, even if the drug does not work. We now know how to do this with appropriate correction, but we didn't always.

Interim Looks

Some people have known about the risks of interim looks, but let me tell you about cimetidine, the first H_2 blocker, approved in 1977

- 4 ulcer healing studies: C vs. placebo
  - 6 week
  - 4 week
  - 2 week X2
- Healing rates were monitored continuously (as each case was completed) and trials were stopped as soon as $p<0.05$; huge inflation of $\propto$ error
- The 2 wk studies worked out. The 4/6 wk studies were stopped but a few more cases wandered in, giving $p>0.05$

To my best knowledge, no one had ever raised the monitoring issue, at least for FDA submitted trials

Interim Looks

Perhaps it was the advent of outcome studies, procedures used in UGDP, BHAT, and growth of DMC's in the 1970's and 1980's but suddenly, by mid 80's or so, all were aware of an inflation and had remedies:

O'Brien-Fleming

Peto

Lan-DeMets, etc.

so everyone now knows you have to 1) correct for multiple looks at data, develop formal stopping rules, and, 2) avoid possible bias, e.g., by making adjustments of endpoints with knowledge of data (which interim efficacy evaluations could lead to), or modifying study design in other ways, such as by changing entry criteria.

BUT, old articles may not deal with this.

Counting All Patients

It seems obvious now, but if, at the end of a study, you can drop out patients for "good" reasons found after the study, you can make any study look favorable.

There were no FDA rules about this until a striking example, the ART (The Anturane Reinfarction Trial) showed us what could happen.

Now, in multiple guidance documents we ask for an accounting of all patients, or at least all patients with data. Any plans to drop anyone need to be specified.

Here's what the ART showed. It was an outcome trial but any study can be manipulated this way, and the omissions generally look very plausible.

Counting All Patients

The Anturane Reinfarction Trial, a study supported in the NEJM by two Dr. Braunwald editorials, seemed to show a survival benefit in post-AMI patients treated with sulfinpyrazone (Anturane), an anti-platelet drug. Our analysis taught us a lot: about cause-specific mortality, multiple endpoints, (unplanned 6 month analysis, unplanned cause-specific mortality analysis), but it was particularly important with respect to dropping patients [Temple R, Pledger G. The FDA's Critique of the Anturane Reinfarction Trial. N Engl J Med 303:1488-1492, 1980]

The Anturane Reinfarction Trial seemed a model effort, one of the first industry-sponsored outcome trials

Features of A.R.T.

Double-Blind (U.A. values hidden) -
Shipped from C-G with numbers.

Randomized in blocks of 10 within each clinic

Placebo-Controlled

Patient Population
- Male or female
- Age 45-70
- AMI 25-35 days before
- ECG Documentation
- Typical Pain History

Enzymes: 2 of CPK, SGOT, LDH had to exceed 2X normal - 72 hr

No cardiomegaly, CHF

>NYHA II, life-limiting disease

Baseline co-variates
- Index MI and later symptoms
- Smoking
- Medications
- Chest x-ray

A.R.T. REPORTED MORTALITY RESULTS

	Pl	S	% ↓ (p)
PATIENTS (Eligible)	783	775	
ALL DEATHS (analyzable)	62	44	29% (p=0.076)
CARDIAC D's	62	43	30.6 ~~32%~~ (p=0.058)
SUDDEN	37	22	43% (p=0.041)
AMI	18	17	--
OTHER	7	4	--
OTHER CV	0	1	--

MORTALITY by CAUSE, TIME

	Pl	S	% ↓ (p-value)
ALL CARDIAC	62	43	30.6% (p=0.058)
ALL CARDIAC			
0-6 M	35	17	50% (p=0.021)
7-24 M	27	26	
SUDDEN			
0-6 M	24	6	74% p=0.003)
7-24 M	13	16	
NON-SUDDEN			
0-6 M	11	11	
7-24 M	14	10	

Ineligible Patients

It was not possible to see this from published reports, but 9 patients <u>who</u> <u>had</u> <u>died</u> were excluded from the results (8 Anturane, one placebo) for being "ineligible" or having poor compliance (pills found in their room). When you put back exclusions, there was no documented effect.

TOTAL CARDIAC DEATHS

	Pl	S
A.R.T.	62	43
POOR COMPLIANCE	1	2
LATE INELIGIBLE	0	6
LESS THAN 7 DAYS	5	4
INELIGIBLE <7D	1	0
TOTAL	69	55
		p=~0.2
LATE DEATHS	13	10
TOTAL	82	65
		p=0.162

Counting All Patients

FDA guidance and Medical Journal Guidance both now clearly call for an accounting of all patients.

It is very tempting to look at data and drop the "outliers," poor compliers, inappropriately entered, etc. It is even plausible. But if not rigorously planned it can be biased and, even if planned, can lead to imbalances that also introduce bias.

Changing Analyses/Multiple Analyses

In the ART, various plausible subanalyses were used, with no real attempt at statistical correction. We saw similar things in DESI. One I recall involved analyses in 2 pain studies

1. The overall studies showed no effect.
2. In study 1, an analysis of moderate and severe patients <u>did</u> show an effect.
3. In study 2, an analysis of mild patients showed an effect.

Subanalysis are possible but must be planned and with appropriate statistical correction.

Active Controls

A longer story than I can discuss here, but showing effectiveness by comparing 2 drugs and seeing “no significant difference,” a once-common approach, is now well-understood to be of little use.

Interpretation of Active Control Trials

Active control equivalence or non-inferiority trials are an intuitively sensible alternative to the placebo-controlled trial, until you realize that effective drugs are not shown effective every time they're studied.

I remember exactly when I realized there was a problem, my epiphany: we saw proposed trials in 1978 or so that were going to compare nadolol with propranolol in angina. But we knew the large majority of placebo-controlled propranolol trials had failed (not shown any effect)

So, how could a finding of no difference between N & P mean anything at all?

It couldn't

Interpretation of Active Control Trials (cont.)

The non-inferiority trial tries to prove effectiveness by showing that the difference between the new drug (T) and the control (C), i.e., C-T, is less than some margin (M), which cannot be greater than the effect you know the control (C) had in this study. (If the difference is larger than all or the effect of C has been lost) But M is not measured (there's no placebo) so it must be assumed, based on past placebo-controlled trial experience. If you show statistically that

$C-T<M$ (97½% CI lower bound)

Then T has some effect > 0

Interpretation of Active Control Trials (cont.)

The critical question is whether this trial could have distinguished the control from placebo and shown an effect of M. If it could have, the trial is said to have "assay sensitivity."

Assay Sensitivity

If a trial <u>has</u> assay sensitivity then if $C-T < M$, T had an effect. If the trial did not have assay sensitivity, then even if $C-T < M$, you have learned nothing

If you don't know whether the trial had assay sensitivity, finding no difference between C and T means either that, in that trial:

Both drugs were effective

Neither drug was effective

Assuring Assay Sensitivity In Non-Inferiority Trials - the Major Problem

In a non-inferiority trial, assay sensitivity <u>is not measured</u> in the trial. That is, the trial itself does not show the study's ability to distinguish active from inactive therapy. Assay sensitivity must, therefore, be deduced or assumed, based on 1) historical experience showing sensitivity to drug effects, 2) a close evaluation of study quality and, particularly important, 3) the similarity of the current trial to trials that were able to distinguish the active control drug from placebo

In many symptomatic conditions, such as depression, pain, allergic rhinitis, IBS, angina, the assumption of assay sensitivity cannot be made, as the following example shows.

TABLE 1. Results (4 week adjusted endpoint Ham-D total scores) of 6 trials comparing a new antidepressant, imipramine, and placebo showing only the new drug vs. imipramine comparison.

Study	Item	Common Baseline	NEW	IMI	"p" two tail	Power to detect 30% difference
R301	HAM-D (n)	23.9	13.4 33	12.8 33	0.78	0.40
G305	HAM-D (n)	26.0	13.0 39	13.4 30	0.86	0.45
C311(1)	HAM-D (n)	28.1	19.4 11	20.3 11	0.81	0.18
V311(2)	HAM-D (n)	29.6	7.3 7	9.5 8	0.63	0.09
F313	HAM-D (n)	37.6	21.9 7	21.9 8	1.0	0.26
K317	HAM-D (n)	26.1	11.2 37	10.8 32	0.85	0.33

TABLE 2. Results (4 week adjusted endpoint Ham-D total scores) of 6 trials comparing a new antidepressant, imipramine, and placebo showing all comparisons.

Study	Item	NEW	IMI	PBO	Baseline HAM-D adjusted	
R301	HAM-D (n)	13.4 33	12.8 33	14.8 36	23.9	
G305	HAM-D (n)	13.0 39	13.4 30	13.9 36	26.0	
C311(1)	HAM-D (n)	19.4 11	20.3 11	18.9 13	28.1	
V311(2)	HAM-D (n)	7.3 7	9.5 8	23.5 7	29.6	*
F313	HAM-D (n)	21.9 7	21.9 8	22 8	37.6	
K317	HAM-D (n)	11.2 37	10.8 32	10.5 36	26.1	

*IMI, NEW vs PBO, "p" less than 0.001

Active Controls

So you can use a non-inferiority design only where you can tell from historical experience that the control drug will almost always have a detectable effect of a defined size in a trial. As noted, few, symptomatic treatments will meet this test.

Number of Studies

As noted, 2 expected but FDAMA (1997) allowed 1 under some circumstances. A Guidance (1998) described cases in which this was reasonable and also addressed the issue of the Quality of evidence, less detached reports, literature, etc.

It described situations in which evidence from other sources (other studies or, sometimes, other drugs or pharmacologic studies, could support one new study of the drug.

One Study Plus Related Studies: Examples

A. Straightforward Cases of "confirmatory evidence" in the form of other adequate and well-controlled studies

1. Studies of different doses, regimens, dosage forms (may need no new study; if needed, generally only one). Anecdote: DESI history, entirely "proof of principle" (different doses, products, dosage forms, regimens, all examined together)

2. Studies in other phases of the same disease. Generally, expect similar direction of response in all stages, though magnitude and B/R may differ (typical in oncology, for same tumor; severities of heart failure)

3. Studies in other populations (if additional studies needed)

4. Combination and Monotherapy; each supports the other (typical in oncology, antihypertensives) - NB - not "automatic;" in one recent case, we did not conclude that an AED effective in combination was shown effective as monotherapy by a single favorable study: the effect was small and needed a larger dose; a second larger and longer study showed no effect.

5. Studies in a closely-related diseases or in pathophysiologically-related conditions: e.g., one study in each of two inflammatory conditions; one study in each of two pain models; anti-platelet drugs in acute coronary syndrome and post-PTCA

One Study Plus Related Studies: Examples

B. More difficult cases

6. Less closely related diseases, similar purpose of therapy. Effectiveness in one tumor might suggest reliance on a single study in a second tumor (possibly depends on tumor types); effectiveness of antibiotic at one site might support another setting with similar pathogens, at least in some sites

7. Studies with 2 different, but related clinical endpoints. Enalapril for CHF supported by one (of 2) exercise tolerance studies and one (dramatic) survival study; given both symptomatic and survival claims. Other examples could include different (but related) tests of depression or cognitive function, effects on survival and recurrent infarction in different studies.

   Issues: Suppose one endpoint is a surrogate; does it support an outcome claim (e.g., lipid-lowering drug with one outcome study and one study showing decreased coronary obstruction). This would seem to depend on amount of support for surrogate and existing outcome data. The surrogate could, of course, be considered "pharmacologic" evidence.

One Study Plus Related Examples

C. Most Difficult Case

8. Support by pharmacologic/pathophysiologic effect

NB: a) this is <u>not</u> the case of whether an accepted surrogate (these lead to ordinary approval) or a "reasonable" surrogate (these lead to accelerated approval), can be used as evidence. They can, although in both cases they generally do not lead to approval of an outcome claim. Could a surrogate be used to support a single study of outcomes?

b) few examples given because this is a treacherous area - there is always <u>some</u> pharmacologic effect; when is it confirmatory?

c) This is not the case where a single persuasive study is sufficient

Principle: "When the pathophysiology of a disease and the mechanism of action of a therapy are well understood, it may be possible to link specific pharmacologic effects to a strong likelihood of clinical effectiveness"

Pharmacologic Effect (cont'd)

Examples cited include:

- Replacement therapy, such as coagulation factor - clear evidence that deficiency leads to disease. Evidence of restoration of the missing physiologic activity provides support

- Correction of inborn effort of metabolism

- Vaccines: one clinical study plus animal challenge protection models, human serological data

- Caveats: Pharmacologic effects have misled (arrhythmia suppression, increased cardiac output by PDE inhibitors)

Pharmacologic Effect (cont'd)

Probably most sensitive case, because of potential broad applicability. Raises critical questions: 1) how much reliance do you place on clinical results with pharmacologically-related drugs; i.e., are the results with those other drugs "confirmatory evidence?" Do we have a "de facto" 1-study standard in this case in general or for serious outcomes? 2) how much weight does belief in mechanism carry; i.e., to what extent is that "relevant science" or "confirmatory evidence?"

Mortality/hospitalization in CHF. ACEI's (several) are effective. Other mechanism adverse

Is one not-overwhelming (but statistically significant) study with ACEI sufficient? Is one study of an angiotensin II inhibitor (probably same mechanism) sufficient? In fact, that has been the standard for ACEI's

Less Detail

Some degree of flexibility is described with respect to our usual level of submitted detail (i.e., <u>everything</u>) but there is clearly expressed concern about journals because their reviewers do not have all the data and peer reviewers are not all equal. But there are strengthening factors; generally some data, such as a protocol and a statistical analysis plan, randomization codes, etc.

Less Detail

Literature <u>can</u> be persuasive; the following increase the "possibility" that we could rely on it

1. Multiple well-designed studies by different investigators
2. Very detailed reports
3. Readily available and appropriate endpoints (not too much judgment)
4. Robust results by a protocol-specified analysis
5. Conducted by groups with track record